# Protein databases

Henrik Nielsen

## Protein databases, historical background

**Swiss-Prot**, http://www.expasy.org/sprot/
Established in 1986 in Swizerland
ExPASy (Expert Protein Analysis System)
Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

**PIR**, http://pir.georgetown.edu/
Established in 1984
National Biomedical Research Foundation, Georgetown University, USA

*In 2002 merged into:*
**UniProt**, http://www.uniprot.org/
A collaboration between SIB, EBI and Georgetown University.


## UniProt

**UniProt Knowledgebase (UniProtKB)**
**UniProt Reference Clusters (UniRef)**
**UniProt Archive (UniParc)**
**UniProt Metagenomic and Environmental Sequence Database (UniMES)**

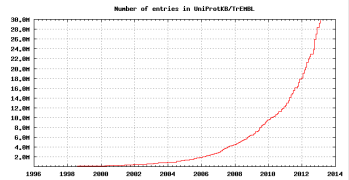
**UniProt Knowledgebase Release 2013_02 (06-Feb-13)** consists of:

**UniProtKB/Swiss-Prot:** Annotated manually (*curated*)
539,165 entries

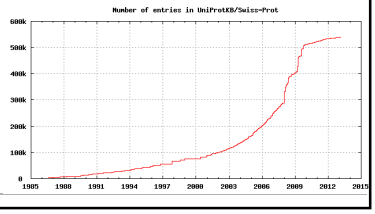
**UniProtKB/TrEMBL:** Computer annotated
29,769,971 entries

## Growth of UniProt

TrEMBL




Swiss-Prot




## Content of UniProt Knowledgebase

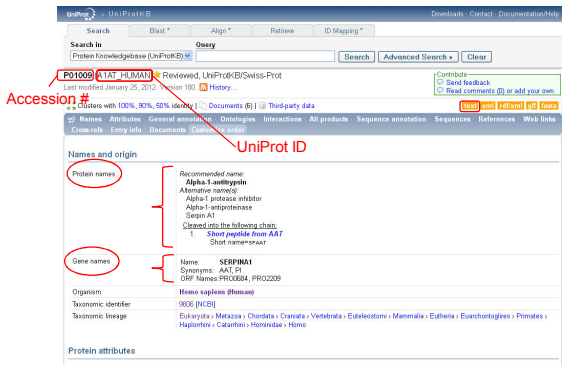
- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

## Amino acid sequences

**From where do you get amino acid sequences?**

- Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
- Direct amino acid sequencing: *Edman degradation*
- Mass spectrometry
- 3D-structures

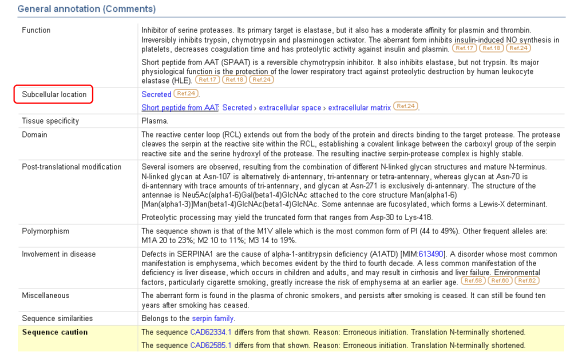
# UniProt entry, formatted view




# UniProt entry, text view (flat file)

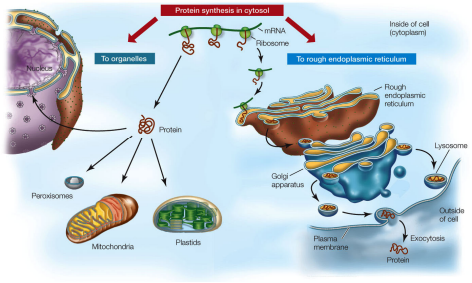
```
ID   A1AT_HUMAN              Reviewed;         418 AA.
AC   P01009; A6PX14; B2RDQ8; Q0PVP5; Q13672; Q53XB8; Q5UUM1; Q7M4R2;
AC   Q86U18; Q86U19; Q96BF9; Q96ES1; Q9P1P0; Q9UCE6; Q9UCM3;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1996, sequence version 3.
DT   25-JAN-2012, entry version 180.
DE   RecName: Full=Alpha-1-antitrypsin;
DE   AltName: Full=Alpha-1 protease inhibitor;
DE   AltName: Full=Alpha-1-antiproteinase;
DE   AltName: Full=Serpin A1;
DE   Contains:
DE     RecName: Full=Short peptide from AAT;
DE              Short=SPAAT;
DE   Flags: Precursor;
GN   Name=SERPINA1; Synonyms=AAT, PI; ORFNames=PRO0684, PRO2209;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX   MEDLINE=84107980; PubMed=6319097;
RA   Bollen A., Herzog A., Cravador A., Herion P., Chuchana P.,
RA   van der Straten A., Loriau R., Jacobs P., van Elsen A.;
RT   "Cloning and expression in Escherichia coli of full-length
RT   complementary DNA coding for human alpha 1-antitrypsin.";
RL   DNA 2:255-264(1983).
...
```

# General annotation (Comments)



# Protein sorting in eukaryotes

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




Different proteins belong to different compartments of the cell – and some belong *outside* the cell

# General annotation (Comments)

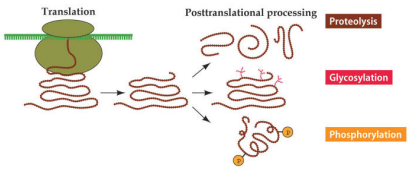
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

**General annotation (Comments)**

| | |
|---|---|
| Function | Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin. (Ref.12) (Ref.19) (Ref.24)<br>Short peptide from AAT (SPAAT) is a reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE). (Ref.17) (Ref.18) (Ref.24) |
| Subcellular location | Secreted (Ref.24).<br>Short peptide from AAT: Secreted › extracellular space › extracellular matrix (Ref.24). |
| Tissue specificity | Plasma. |
| Domain | The reactive center loop (RCL) extends out from the body of the protein and directs binding to the target protease. The protease cleaves the serpin at the reactive site within the RCL, establishing a covalent linkage between the carboxyl group of the serpin reactive site and the serine hydroxyl of the protease. The resulting inactive serpin-protease complex is highly stable. |
| Post-translational modification | Several isomers are observed, resulting from the combination of different N-linked glycan structures and mature N-terminus. N-linked glycan at Asn-107 is alternatively di-antennary, tri-antennary or tetra-antennary, whereas glycan at Asn-70 is di-antennary with trace amounts of tri-antennary, and glycan at Asn-271 is exclusively di-antennary. The structure of the antennae is Neu5Ac(alpha1-6)Gal(beta1-4)GlcNAc attached to the core structure Man(alpha1-6)[Man(alpha1-3)]Man(beta1-4)GlcNAc(beta1-4)GlcNAc. Some antennae are fucosylated, which forms a Lewis-X determinant.<br>Proteolytic processing may yield the truncated form that ranges from Asp-30 to Lys-418. |
| Polymorphism | The sequence shown is that of the M1V allele which is the most common form of PI (44 to 49%). Other frequent alleles are: M1A 20 to 23%; M2 10 to 11%; M3 14 to 19%. |
| Involvement in disease | Defects in SERPINA1 are the cause of alpha-1-antitrypsin deficiency (A1ATD) [MIM:613490]. A disorder whose most common manifestation is emphysema, which becomes evident by the third to fourth decade. A less common manifestation of the deficiency is liver disease, which occurs in children and adults, and may result in cirrhosis and liver failure. Environmental factors, particularly cigarette smoking, greatly increase the risk of emphysema at an earlier age. (Ref.59) (Ref.60) (Ref.62) |
| Miscellaneous | The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased. |
| Sequence similarities | Belongs to the serpin family. |
| **Sequence caution** | The sequence CAD62334.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened.<br>The sequence CAD62585.1 differs from that shown. Reason: Erroneous initiation. Translation N-terminally shortened. |

# Post-translational modifications

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




Many proteins are *modified* after their synthesis in order to become active

**Proteolysis:** Cleavage of *signal peptides*, *propeptides* or *initiator methionine*

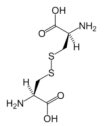
**Glycosylation:** Particularly common in proteins on the *surface* of cells. Also plays a role in sorting of proteins til *lysosomes*

**Phosphorylation:** Often *reversible*. Regulates the *activity* of many enzymes

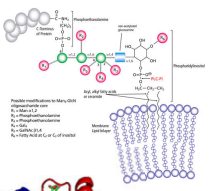
# More post-translational modifications

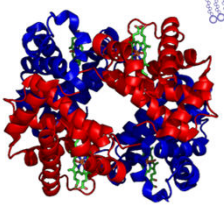
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Lipid anchors
  - (e.g. GPI anchors)
- Disulfide bonds



- Prosthetic groups
  - (*e.g.* metal ions)





# General annotation (Ontologies)

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Ontologies

**Keywords**

| | |
|---|---|
| Biological process | Acute phase<br>Blood coagulation |
| Cellular component | Extracellular matrix<br>Secreted |
| Coding sequence diversity | Alternative splicing<br>Polymorphism |
| Domain | Signal |
| Molecular function | Protease inhibitor<br>Serine protease inhibitor |
| PTM | Glycoprotein |
| Technical term | 3D-structure<br>Complete proteome<br>Direct protein sequencing<br>Reference proteome |

**Gene Ontology (GO)**

| | |
|---|---|
| Biological process | acute-phase response<br>Inferred from electronic annotation. Source: UniProtKB-KW<br>platelet activation<br>Traceable author statement. Source: Reactome<br>platelet degranulation<br>Traceable author statement. Source: Reactome<br>regulation of proteolysis<br>Inferred from Biological aspect of Ancestor. Source: RefGenome |
| Cellular component | extracellular space<br>Inferred from mutant phenotype. Source: UniProtKB<br>platelet alpha granule lumen<br>Traceable author statement. Source: Reactome<br>proteinaceous extracellular matrix<br>Inferred from electronic annotation. Source: UniProtKB-SubCell |
| Molecular function | protease binding<br>Inferred from physical interaction. Source: UniProtKB<br>serine-type endopeptidase inhibitor activity<br>Non-traceable author statement. Source: UniProtKB |


QuickGO - http://www.ebi.ac.uk/QuickGO

# Sequence annotation (Feature Table)

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Sequence annotation (Features)

| Feature key | Position(s) | Length | Description | Graphical view | Feature identifier |
|---|---|---|---|---|---|
| **Molecule processing** | | | | | |
| Signal peptide | 1 – 24 | 24 | | | |
| Chain | 25 – 418 | 394 | Alpha-1-antitrypsin | | PRO_0000032377 |
| Peptide | 375 – 418 | 44 | Short peptide from AAT | | PRO_0000364030 |
| **Regions** | | | | | |
| Region | 368 – 392 | 25 | RCL | | |
| **Sites** | | | | | |
| Site | 382 – 383 | 2 | Reactive bond | | |
| **Amino acid modifications** | | | | | |
| Modified residue | 256 | 1 | S-cysteinyl cysteine | | |
| Glycosylation | 70 | 1 | N-linked (GlcNAc...) (complex) | | |
| Glycosylation | 107 | 1 | N-linked (GlcNAc...) (complex) | | |
| Glycosylation | 271 | 1 | N-linked (GlcNAc...) (complex) | | |
| **Natural variations** | | | | | |
| Alternative sequence | 397 – 418 | 112 | Missing in isoform 3. | | VSP_028890 |
| Alternative sequence | 366 – 418 | 63 | AVHKA...NPTQK → VRSP in isoform 2. | | VSP_028889 |
| Natural variant | 4 | 1 | S → L in Z-Wrexham. | | VAR_006978 |
| Natural variant | 26 | 1 | D → A in V-Munich. | | VAR_006979 |
| Natural variant | 37 | 1 | T → A.<br>Corresponds to variant rs11552062 [ dbSNP \| Ensembl ]. | | VAR_051938 |
| Natural variant | 58 | 1 | A → T in MS-Karlsruhe. | | VAR_006980 |
| Natural variant | 63 | 1 | R → C in I. | | VAR_006981 |
| | | | Corresponds to variant rs28931570 [ dbSNP \| Ensembl ]. | | |

# Secondary structure (Feature Table)

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

**Secondary structure**

Helix Strand Turn

Details...

| | | |
|---|---|---|
| Turn | 48 – 50 | 3 |
| Helix | 51 – 68 | 18 |
| Beta strand | 70 – 72 | 3 |
| Beta strand | 74 – 76 | 3 |
| Helix | 78 – 89 | 12 |
| Helix | 94 – 103 | 10 |
| Turn | 108 – 110 | 3 |
| Helix | 113 – 127 | 15 |
| Beta strand | 135 – 145 | 11 |
| Helix | 152 – 160 | 9 |
| Beta strand | 165 – 169 | 5 |
| Helix | 174 – 188 | 15 |
| Turn | 189 – 191 | 3 |
| Beta strand | 206 – 215 | 10 |
| Beta strand | 218 – 220 | 3 |
| Helix | 224 – 226 | 3 |
| Beta strand | 228 – 237 | 10 |
| Beta strand | 239 – 256 | 18 |
| Turn | 257 – 260 | 4 |
| Beta strand | 261 – 268 | 8 |
| Turn | 269 – 271 | 3 |
| Beta strand | 272 – 279 | 8 |

# Protein structure

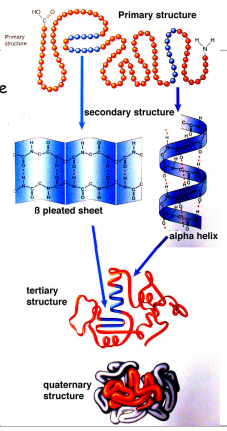
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Primary structure: Amino acid sequence

Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

Quaternary structure: subunits




# Evidence (Comments, Feature Table)

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

**Q43495** (108_SOLLC) Reviewed, UniProtKB/Swiss-Prot

Last modified March 2, 2010. Version 49. History...

**General annotation (Comments)**

| | |
|---|---|
| Subcellular location | Secreted (Potential) |
| Tissue specificity | Stamen- and tapetum-specific. |
| Sequence similarities | Belongs to the A9/FIL1 family. |

**Sequence annotation (Features)**

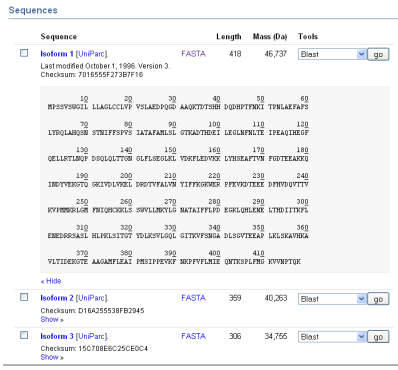
| | Feature key | Position(s) | Length | Description | Graphical view | Feature identifier |
|---|---|---|---|---|---|---|
| **Molecule processing** | | | | | | |
| | Signal peptide | 1 – 30 | 30 | (Potential) | | |
| | Chain | 31 – 102 | 72 | Protein 108 | | PRO_0000000238 |
| **Amino acid modifications** | | | | | | |
| | Disulfide bond | 41 ↔ 77 | | (By similarity) | | |
| | Disulfide bond | 51 ↔ 66 | | (By similarity) | | |
| | Disulfide bond | 67 ↔ 92 | | (By similarity) | | |
| | Disulfide bond | 79 ↔ 99 | | (By similarity) | | |

## Evidence/Confidence types

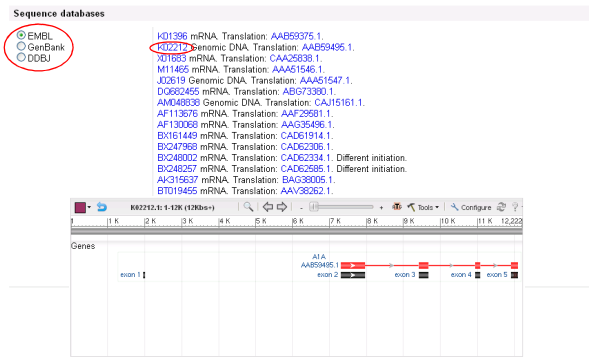
3 types of *non-experimental qualifiers* in Sequence annotation and General comment:

- *Potential:* Predicted using sequence analysis
- *Probable:* Uncertain experimental evidence
- *By similarity:* Predicted using sequence similarity

## UniProt entry, sequence(s)




## Cross-references, nucleotide sequences



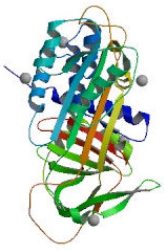
## Cross-references, 3D structure

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

**3D structure databases**

- (•) PDBe
- ( ) RCSB PDB
- ( ) PDBj

| Entry | Method | Resolution (Å) | Chain | Positions | PDBsum |
|---|---|---|---|---|---|
| 1ATU | X-ray | 2.70 | A | 45-418 | [»] |
| 1D5S | X-ray | 3.00 | A | 44-377 | [»] |
| | | | B | 378-418 | [»] |
| 1EZX | X-ray | 2.60 | A | 48-382 | [»] |
| | | | B | 383-418 | [»] |
| 1HP7 | X-ray | 2.10 | A | 25-418 | [»] |
| 1IZ2 | X-ray | 2.20 | A | 25-418 | [»] |
| 1WCT | X-ray | 3.46 | A | 25-418 | [»] |
| 1O08 | X-ray | 2.65 | A | 26-418 | [»] |
| 1OPH | X-ray | 2.30 | A | 26-418 | [»] |
| 1PSI | X-ray | 2.92 | A | 26-418 | [»] |
| 1QLP | X-ray | 2.00 | A | 26-418 | [»] |
| 1QMB | X-ray | 2.60 | A | 49-376 | [»] |
| | | | B | 377-418 | [»] |
| 2D26 | X-ray | 3.30 | A | 26-382 | [»] |
| | | | B | 383-418 | [»] |
| 2QUG | X-ray | 2.00 | A | 25-418 | [»] |
| 3CWL | X-ray | 2.44 | A | 25-418 | [»] |
| 3CWM | X-ray | 2.51 | A | 25-418 | [»] |
| 3DRM | X-ray | 2.20 | A | 26-418 | [»] |
| 3DRU | X-ray | 3.20 | A/B/C | 26-418 | [»] |
| 3NDD | X-ray | 1.50 | A | 46-372 | [»] |
| | | | B | 383-418 | [»] |
| 3NDF | X-ray | 2.70 | A | 46-381 | [»] |
| | | | B | 383-418 | [»] |
| 3T1P | X-ray | 3.90 | A | 48-418 | [»] |
| 7API | X-ray | 3.00 | A | 36-382 | [»] |
| | | | B | 383-418 | [»] |
| 8API | X-ray | 3.10 | A | 36-382 | [»] |
| | | | B | 383-418 | [»] |
| 9API | X-ray | 3.00 | A | 36-382 | [»] |
| | | | B | 383-418 | [»] |



## Cross-references

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Other databases linked from UniProt
(there are ~100 in total):

- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gene expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

# Translation and Reading Frames

# The genetic code



| First letter | Second letter: U | Second letter: C | Second letter: A | Second letter: G | Third letter |
|---|---|---|---|---|---|
| U | UUU, UUC: Phenylalanine<br>UUA, UUG: Leucine | UCU, UCC, UCA, UCG: Serine | UAU, UAC: Tyrosine<br>UAA: Stop codon<br>UAG: Stop codon | UGU, UGC: Cysteine<br>UGA: Stop codon<br>UGG: Tryptophan | U, C, A, G |
| C | CUU, CUC, CUA, CUG: Leucine | CCU, CCC, CCA, CCG: Proline | CAU, CAC: Histidine<br>CAA, CAG: Glutamine | CGU, CGC, CGA, CGG: Arginine | U, C, A, G |
| A | AUU, AUC, AUA: Isoleucine<br>AUG: Methionine; start codon | ACU, ACC, ACA, ACG: Threonine | AAU, AAC: Asparagine<br>AAA, AAG: Lysine | AGU, AGC: Serine<br>AGA, AGG: Arginine | U, C, A, G |
| G | GUU, GUC, GUA, GUG: Valine | GCU, GCC, GCA, GCG: Alanine | GAU, GAC: Aspartic acid<br>GAA, GAG: Glutamic acid | GGU, GGC, GGA, GGG: Glycine | U, C, A, G |

- Degenerate (*redundant*) but not ambiguous
- *Almost* universal (deviations found in mitochondria)

# Reading Frames 1



A piece of an mRNA-strand:

```
5'                                                  3'
augcccaagcugaauagcguagagggguuuucaucauuugaggacgauguauaa
```

can be divided into triplets (*codons*) in three ways:

```
1 aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  ugc cca agc uga aua gcg uag agg ggu uuu cau cau uug agg acg aug uau
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gcu gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Each possible set of triplets is called a *reading frame*.

# Reading Frames 2



Since there are two strands in DNA, there are *six* possible reading frames in a piece of DNA (three in each direction):

```
3    A  Q  A  E  *  R  R  G  V  F  I  I  *  G  R  C  I
2   C  P  S  *  I  A  *  R  G  F  H  H  L  R  T  M  Y
1  M  P  K  L  N  S  V  E  G  F  S  S  F  E  D  D  V  *
5' ATGCCCAAGCTGAATAGCGTAGAGGGGTTTTCATCATTTGAGGACGATGTATAA 3'
3' TACGGGTTCGACTTATCGCATCTCCCCAAAAGTAGTAAACTCCTGCTACATATT 5'
    H  G  L  Q  I  A  Y  L  P  K  *  *  K  L  V  I  Y  L  -1
     G  L  S  F  L  T  S  P  N  E  D  N  S  S  S  T  Y  -2
    A  W  A  S  Y  R  L  P  T  K  M  M  Q  P  R  H  I   -3
```

A reading frame from a start codon to the first stop codon is called an *open* reading frame (underlined above).

# Virtual Ribosome – required reading!

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS **CBS**

*Nucleic Acids Research, 2006, Vol. 34, Web Server issue* **W385–W388**
*doi:10.1093/nar/gkl252*

## Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation

**Rasmus Wernersson***

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received February 14, 2006; Revised March 1, 2006; Accepted March 20, 2006

### ABSTRACT

**Virtual Ribosome is a DNA translation tool with two areas of focus. (i) Providing a strong translation tool in its own right, with an integrated ORF finder, full support for the IUPAC degenerate DNA alphabet and all translation tables defined by the NCBI taxonomy group, including the use of alternative start codons. (ii) Integration of sequences feature annotation—in particular, native support for working with files containing intron/exon structure annotation. The software is available for both download and online use at http://www.cbs.dtu.dk/services/VirtualRibosome/.**

### INTRODUCTION

A large number of software packages for translating DNA sequences already exist, as services on the World Wide

This makes it easy to build datasets that can be used for analyzing how the underlying exon structure is reflected in the protein [e.g. how exon modules maps onto the 3D structure of the protein, see the FeatureMap3D server (4) elsewhere in this issue].

### SOFTWARE FEATURES

#### Support for the degenerate nucleotide alphabet

The software has full support for the IUPAC alphabet (Table 1) for degenerate nucleotides. For example, the codon TCN correctly translates to S (serine) and not X (unknown) as often seen in other translators.

#### Support for a wide range of translation tables

Full support for all translation tables defined by the NCBI taxonomy group (5) (see the list below). The command-line version of the software also has support for